

1 Introduction

1.1 Background

This entry examines the distributions of time to the initial response in milliseconds across a series of choice tasks. The R code shows how to illustrate these times as histograms, tables, and line plots as well as how to calculate their medians and interquartile ranges (IQR).

For this worked example, data on initial response times was extracted from the 2016 predictive modeling competition in HPR (Jakubczyk, Craig, et al. 2017). In 2016, 4088 US participants responded to 20 paired comparisons, choosing between two alternative health outcomes. Apart from initial response time, respondents may change their answers before proceeding to the next task (time to last response) or may spend additional time on the page before proceeding to the next task (i.e., page time); nevertheless, the time to initial response is a common behavioral measure of task difficulty in HPR. For this analysis, initial response times were truncated at five minutes (300,000 milliseconds).

We will continue to examine this topic. Please see the entry *Correlations between initial response times*, where we examine the variance, co-variance, correlations of time to the initial response.

1.2 Load libraries and source files

Notes: Change the **working directory** to the location of the source files on your computer. To do so, you need to replace the inside of `setwd()` with the location of data file on your computer.

```
setwd("C:\\Users\\aaa\\OneDrive\\USF\\Dr. Craig")
library(knitr) #this is for the function "kable," which makes well-organized tables.
library(tidyverse) #this is for the function "read_csv."
library(tinytex)
library(gt)
data1 <- read_csv("resp1wave1_220723.csv")
```

2 Visualize distributions of initial response times

2.1 Background

- *Definition:* A **histogram** is a diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the range divided by the number of bins. To create a histogram of initial response times, you must decide on the number of bins, K , and the range of histogram, $[r_0, r_K)$. When you decide the number of bins, K , “Sturges’ rule” might give you a clue. Note that this rule just gives you a rough indication. (This entry does not follow this rule.)

$$\text{Sturges' rule: } K = \log_2 (N \times T) + 1$$

. You must decide r_0 and r_K .

$$r_0, r_K \text{ s.t. } \min(x_{it}), \max(x_{it}) \in [r_0, r_K)$$

.

Let x_{it} be the initial response time for respondent i and task t ($1 \leq i \leq N$; $1 \leq t \leq T$). A histogram needs to satisfy Equation (1).

$$N \times T = \sum_{k=1}^K m_k \quad (1)$$

where

N = the total number of respondents

T = the total number of tasks per respondent

K = the total number of bins ($1 \leq K \leq N \times T$)

m_k = the number of x_{it} within the k^{th} bin ($1 \leq k \leq K$)

= the number of x_{it} ($r_{k-1} \leq x_{it} < r_k$)

= frequency with the k^{th} bin

r_j = the upper boundary of the j^{th} bin ($\min(k) - 1 \leq j \leq \max(k)$, that is, $0 \leq j \leq K$)

= the lower boundary of the $j + 1^{th}$ bin

$r_j - r_{j-1}$ = the width of each bin

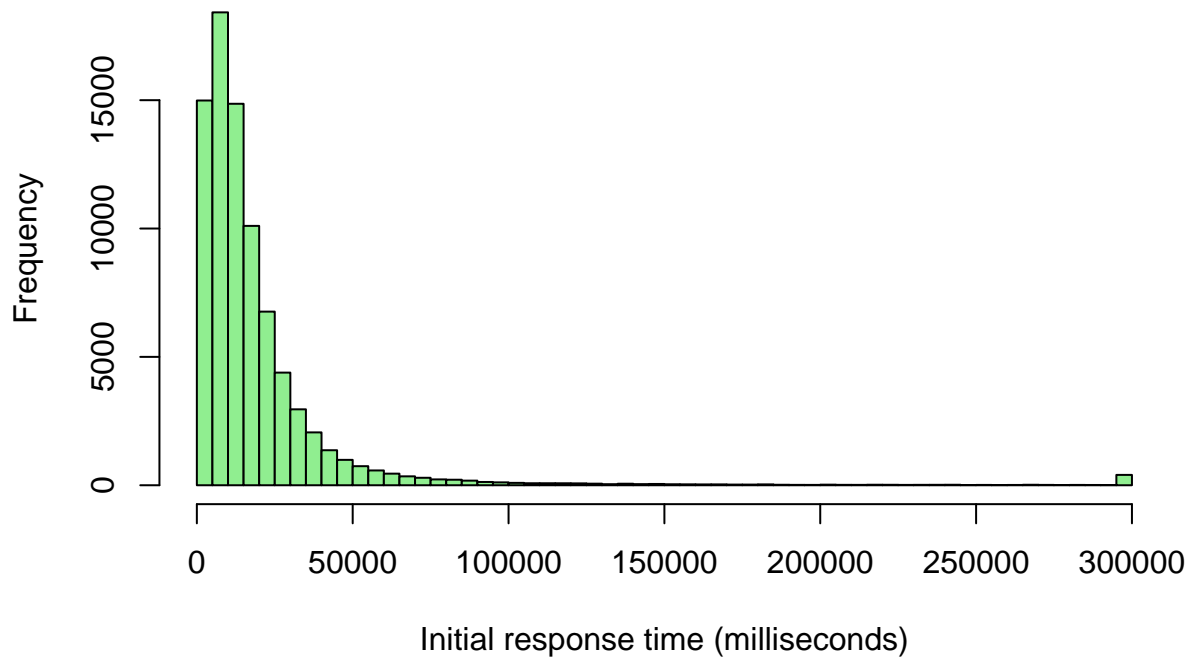
$$= \frac{\max(r_j) - \min(r_j)}{K} = \frac{r_K - r_0}{K}$$

2.2 Example I

Now, let $K = 50$ (50 bins or breaks). This histogram represents the distribution of the initial response times x_{it} for all task ($N = 4088$; $T = 20$).

```
hist(data1$time,
     breaks = 50,
     main = "Histogram 2.2: Initial response time for the first task",
     xlab = "Initial response time (milliseconds)",
     col = "lightgreen")
```

Histogram 2.2: Initial response time for the first task

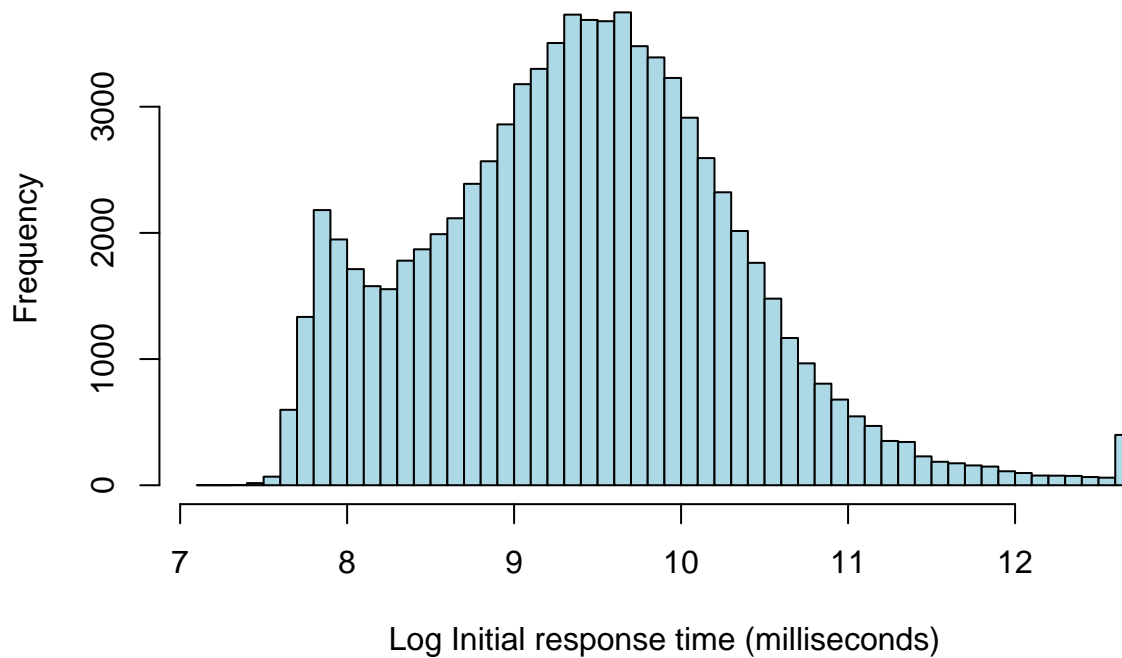


2.3 Example II

Now, replace $x_{it} = \log(x_{it})$. This histogram represent the distribution of log initial response times for all tasks ($N = 4088$; $T = 20$)

```
hist(log(data1$time),  
     breaks = 50,  
     main = "Histogram 2.3: Log Initial response time for the first task",  
     xlab = "Log Initial response time (milliseconds)",  
     col = "lightblue")
```

Histogram 2.3: Log Initial response time for the first task



The histogram better illustrates that the distribution is bimodal (i.e., two local maxima).

2.4 Example III

This code produces overlapping histograms to compare the distribution of log initial response times between the first and last tasks ($N = 4088$; $t = 1, 20$)

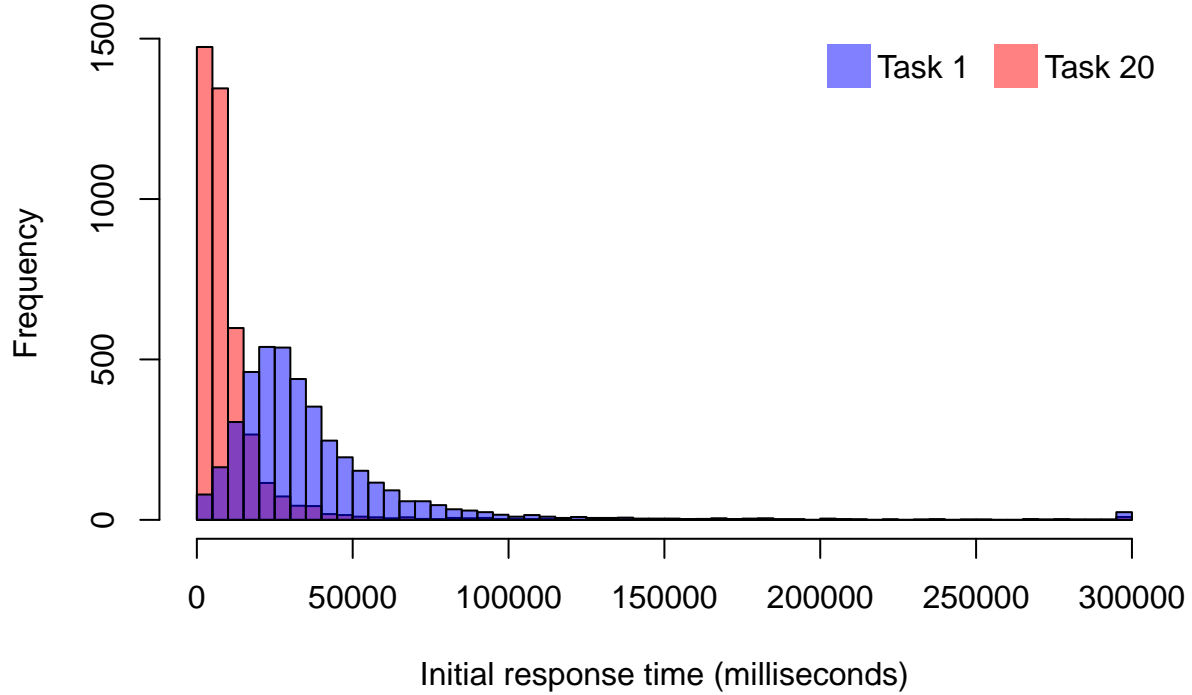
```
d2 = matrix()
for(i in 1:20) {d1 <- data1[data1$task == i, "time"]
d2 = cbind(d2, d1)
}
d3 <- as.data.frame(d2[, -1])
colnames(d3) = paste0("Task", 1:max(data1$task))
rownames(d3) = paste0("ID ", 1:max(data1$survey_id))
xh <- sort(d3$Task1) %>% data.frame()#this is task1 sorted by the value (increasing).
colnames(xh) = c("task1")

hist(d3[,20], breaks = 50, col="#FF00007F",
     main="Histogram 2.4: Initial response time for the first and last tasks",
     xlab="Initial response time (milliseconds)")
hist(d3[,1], breaks = 50, col="#0000FF7F", add=T)

legend("topright",
      legend=c("Task 1", "Task 20"),
      pch=15,
```

```
col=c("#0000FF7F", "#FF00007F"),
bty="n",
ncol=2,
pt.cex=3
)
```

Histogram 2.4: Initial response time for the first and last tasks



The overlay histogram shows distributional differences between the first and last tasks.

3 Visualize the percentiles of initial response times

3.1 Background

- *Definition* : The k^{th} **percentile** is a number such that k percent of observations have an equal or smaller value than that number. For example, if the 25th percentile is 1256, then 25 percent of observations are equal or smaller than 1256 and 75 percent are larger.
- *Definition*: The **median** is the 50th percentile (Q2).
Assume initial response times for task t , x_t were sorted from smallest to largest, $x_t[1] \leq x_t[2] \leq \dots \leq x_t[N]$.
When N is even, that is, $\exists a \in \mathbb{N}$ s.t. $N = 2a$, $median(Q2) = \frac{1}{2}(x_t[\frac{N}{2}] + x_t[\frac{N}{2} + 1])$.
When N is odd, that is, $\exists b \in \mathbb{N}$ s.t. $N = 2b + 1$, $median(Q2) = x_t[\frac{N+1}{2}]$.
- *Definition*: The **interquartile range** (IQR) is the distance between the 25th percentile (Q1) and the 75th percentile (Q3). That is, $IQR = Q3 - Q1$.

Q1 and Q3 could be given by Equation (2) and (3), respectively. However, depending on the value of N , $\frac{N+1}{4}$ and $\frac{3(N+1)}{4}$ could be non-integers. In those cases, you need to apply a linear interpolation to $x_t[\lceil \frac{N+1}{4} \rceil]$ and $x_t[\lceil \frac{3(N+1)}{4} \rceil]$ for Q1 and $x_t[\lfloor \frac{3(N+1)}{4} \rfloor]$ and $x_t[\lfloor \frac{3(N+1)}{4} \rfloor]$ for Q3 so that you are able to calculate them (shown in cases (i)-(iv)).

$$Q1 = x_t[\frac{N+1}{4}] \quad (2)$$

$$Q3 = x_t[\frac{3(N+1)}{4}] \quad (3)$$

The way of linear interpolation are divided into four cases (i)-(iv) following Equation (4).

$$\forall N \in \mathbb{N}, \exists h \in \mathbb{N} \text{ s.t. } N = \begin{cases} 4h \\ 4h - 1 \\ 4h + 1 \\ 4h + 2 \end{cases} \quad (4)$$

(i) $N = 4h$

$$Q1 = \frac{3}{4}x_t[h] + \frac{1}{4}x_t[h+1]$$

$$Q3 = \frac{1}{4}x_t[3h] + \frac{3}{4}x_t[3h+1]$$

$$IQR = Q3 - Q1$$

$$= (\frac{1}{4}x_t[3h] + \frac{3}{4}x_t[3h+1]) - (\frac{3}{4}x_t[h] + \frac{1}{4}x_t[h+1])$$

(ii) $N = 4h - 1$

$$Q1 = \frac{1}{2}x_t[h] + \frac{1}{2}x_t[h+1]$$

$$Q3 = \frac{1}{2}x_t[3h+1] + \frac{1}{2}x_t[3h+2]$$

$$IQR = Q3 - Q1$$

$$= (\frac{1}{2}x_t[3h+1] + \frac{1}{2}x_t[3h+2]) - (\frac{1}{2}x_t[h] + \frac{1}{2}x_t[h+1])$$

(iii) $N = 4h + 1$

$$Q1 = x_t[h]$$

$$Q3 = x_t[3h]$$

$$IQR = Q3 - Q1$$

$$= x_t[3h] - x_t[h]$$

(iv) $N = 4h + 2$

$$Q1 = \frac{1}{4}x_t[h] + \frac{3}{4}x_t[h+1]$$

$$Q3 = \frac{3}{4}x_t[3h+2] + \frac{1}{4}x_t[3h+3]$$

$$IQR = Q3 - Q1$$

$$= (\frac{3}{4}x_t[3h+2] + \frac{1}{4}x_t[3h+3]) - (\frac{1}{4}x_t[h] + \frac{3}{4}x_t[h+1])$$

3.2 Example I

This code calculates percentiles of initial response times for the first task ($t = 1$) using the sorted data $x_{i1}[h]$ as well as the command `quantile()` for comparison. Table 3.2.1 shows Q1, Q2, and Q3 for the first task ($t = 1$) that are calculated by the formulas listed above. The command `quantile()` has 9 different types of algorithms (default is `type = 7`). Table 3.2.2 shows percentiles of initial response times for the first task ($t = 1$) by 9 different algorithms. The way of calculation explained above is “`type = 6`”. We have the same calculation results between “`type = 6`” in Table 3.2.2 and Table 3.2.1.

```
h <- floor(nrow(xh)/4)
Q1 <- 0.75*xh[h,1]+0.25*xh[h+1,1]
Q3 <- 0.25*xh[3*h,1]+0.75*xh[3*h+1,1]
Q2 <- 0.5*xh[2*h,1]+0.5*xh[2*h+1,1]
IQRf <- data.frame(Q1=Q1,Q2=Q2,Q3=Q3)
```

```
kable(IQRf, caption="Table 3.2.1: Percentiles by formula for the first task (t=1)",)
```

Table 3.2.1: Percentiles by formula for the first task (t=1)

Q1	Q2	Q3
20148	29645	43792.75

```
IQR1<- NULL
for(i in 1:9){
  IQR0 <- quantile(xh[,1], type=i)
  IQR1 <-rbind(IQR1, IQR0)
}#IQR1 stores 9 types of IQR for task1. Type = 6 is what the author showed in formulas.
rownames(IQR1) = paste0("IQR type=", 1:9)
colnames(IQR1) = c("1st percentile", "25th percentile", "50th percentile",
                  "75th percentile", "100th percentile")
kable(IQR1, caption="Table 3.2.2: 9 types of percentiles for the first task (t=1)",)
```

Table 3.2.2: 9 types of percentiles for the first task (t=1)

	1st percentile	25th percentile	50th percentile	75th percentile	100th percentile
IQR type=1	2023	20146.00	29644	43792.00	3e+05
IQR type=2	2023	20150.00	29645	43792.50	3e+05
IQR type=3	2023	20146.00	29644	43792.00	3e+05
IQR type=4	2023	20146.00	29644	43792.00	3e+05
IQR type=5	2023	20150.00	29645	43792.50	3e+05
IQR type=6	2023	20148.00	29645	43792.75	3e+05
IQR type=7	2023	20152.00	29645	43792.25	3e+05

	1st percentile	25th percentile	50th percentile	75th percentile	100th percentile
IQR type=8	2023	20149.33	29645	43792.58	3e+05
IQR type=9	2023	20149.50	29645	43792.56	3e+05

Again, this example suggests that the commands produce the same results as the sorting method.

3.3 Example II

This code calculates the same results for each of the twenty tasks and visualizes the results using a table.

```
m2 =NULL
ir2 = NULL
iqr2 = NULL
for(i in 1:20) {d1 <- data1[data1$task == i, "time"]
m0 = median(d1$time)
m2 = rbind(m2, m0)
ir0 = quantile(d1$time, type=6)
ir2 = rbind(ir2, ir0)
iqr0 = IQR(d1$time)
iqr2 = rbind(iqr2, iqr0)
table1 = cbind(m2,iqr2, ir2)
}
rownames(table1) = paste0("task ", 1:20)
colnames(table1) <- c("Median", "IQR", "1st percentile", "25th percentile",
"50th percentile", "75th percentile ", "100th percentile")
table3.3 <- data.frame(cbind(c(1:20),table1))
colnames(table3.3) <- c("t th task", "Median", "IQR", "1st percentile", "25th percentile",
"50th percentile", "75th percentile ", "100th percentile")
kable(table3.3, align = "c", caption = "***Table 3.3: Median and IQR**",
row.names = FALSE, escape = FALSE, centering = T)
```

Table 3.3: Median and IQR

t th task	Median	IQR	1st percentile	25th percentile	50th percentile	75th percentile	100th percentile
1	29645.0	23640.25	2023	20148.00	29645.0	43792.75	3e+05
2	20553.0	18588.75	2050	12989.75	20553.0	31580.00	3e+05
3	19209.5	17642.50	1475	11748.25	19209.5	29391.75	3e+05
4	17394.5	16838.00	1590	10981.00	17394.5	27827.00	3e+05
5	16897.5	16081.25	1998	10357.25	16897.5	26443.00	3e+05
6	16023.5	15845.00	1732	9771.75	16023.5	25618.75	3e+05
7	15416.5	14887.75	1787	9422.50	15416.5	24325.75	3e+05
8	15081.0	15779.50	1700	9045.00	15081.0	24827.50	3e+05
9	14317.0	14732.75	1786	8564.25	14317.0	23308.50	3e+05
10	14210.5	14608.50	1669	8305.25	14210.5	22914.75	3e+05
11	13901.0	13757.50	1723	8545.25	13901.0	22321.75	3e+05
12	9884.0	10562.50	1938	5845.25	9884.0	16412.75	3e+05
13	8982.0	9519.50	1818	5105.75	8982.0	14630.25	3e+05

t th task	Median	IQR	1st percentile	25th percentile	50th percentile	75th percentile	100th percentile
14	8580.0	9704.75	1944	4811.75	8580.0	14526.00	3e+05
15	8158.5	9037.50	1829	4459.25	8158.5	13505.75	3e+05
16	7905.5	8648.75	1801	4377.00	7905.5	13033.25	3e+05
17	7386.0	8255.75	1847	4080.75	7386.0	12338.00	3e+05
18	7250.5	8094.75	1538	4111.00	7250.5	12211.25	3e+05
19	7133.5	7991.25	1290	4042.75	7133.5	12036.50	3e+05
20	6908.5	7638.25	1731	3904.50	6908.5	11545.25	3e+05

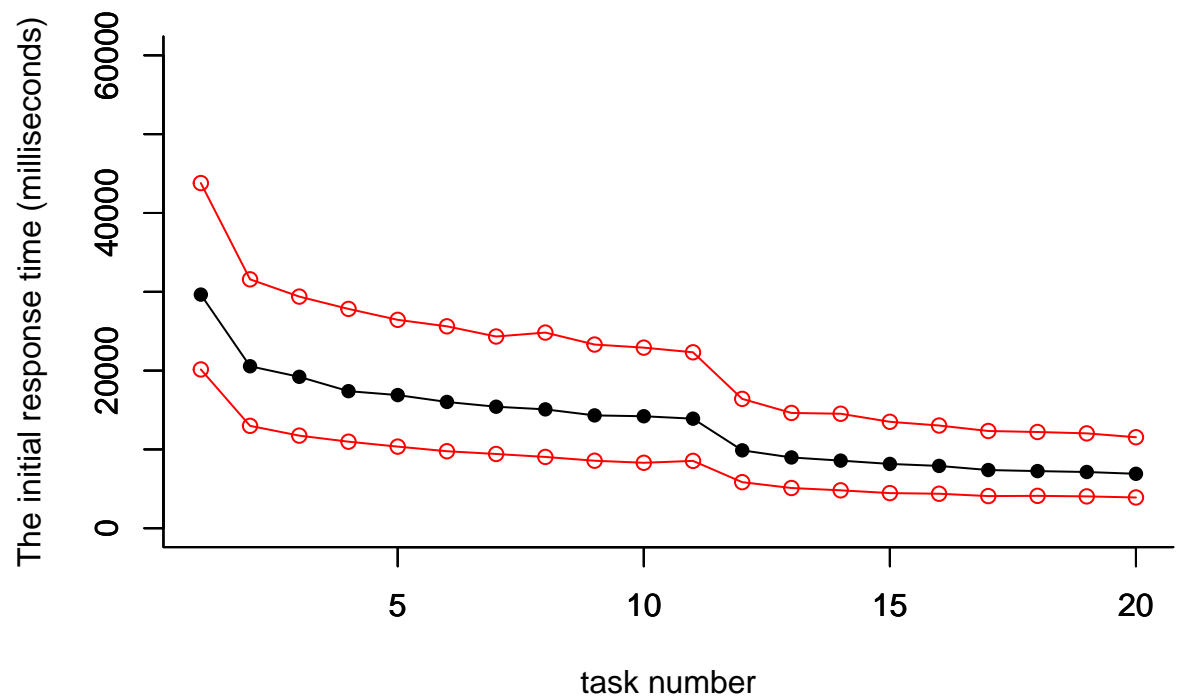
The largest changes in initial response times occurs between the first and second task and between the eleventh and twelfth task.

3.4 Example III

This code visualizes the results taken from Table 3.3 using a line plot, where the black dots represent the median (Q2) and the red dots represent the 25th and 75th percentiles (Q1 and Q3).

```
xmax <- 20
xmin <- 1
ymax <- 60000
ymin <- 0
plot(table3.3['t th task', table3.3$Median, bty = "l", pch = 16, type = "o",
      xlim = c(xmin, xmax), ylim = c(ymin, ymax),
      xlab = NA, ylab = NA, )
par(new=T)
plot(table3.3['t th task', table3.3$'25th percentile', bty = "l", pch = 1,
      col = "red", type = "o",
      xlim = c(xmin, xmax), ylim = c(ymin, ymax), xlab = NA, ylab = NA, )
par(new=T)
plot(table3.3['t th task', table3.3$'75th percentile', bty = "l", pch = 1,
      col = "red", type = "o",
      xlim = c(xmin, xmax), ylim = c(ymin, ymax),
      xlab = "task number", ylab = "The initial response time (milliseconds)",
      main = "Graph 3.4: The median and interquartile range of the initial response time")
```

Graph 3.4: The median and interquartile range of the initial response t



Generally the 25th, 50th, and 75th percentiles decrease monotonically by task and the IQR decreases monotonically by task.

4 Reference

Jakubczyk, M., Craig, B. M., Barra, M., Groothuis-Oudshoorn, C. G. M., Hartman, J. D., Huynh, E., Ramos-Goñi, J. M., Stolk, E. A., & Rand, K. (2017). Choice defines value: A predictive modeling competition in Health Preference Research. *Value in Health*, 21(2), 229–238. <https://doi.org/10.1016/j.jval.2017.09.016>

5 How to cite this entry

Okubo, S. (yyyy, month dd). Illustrating time to initial response in choice tasks. *R4HPR*. <https://r4hpr.org/visor/?e=analysis-of-initial-response-time-across-20-pair-comparisons>